Depression Detection in Social Media Posts

Theodor Cucu **Rujul Gandhi**

Abstract

Depression is a very common mental illness which also happens to be subtle in its manifestation. Natural Language Processing can be used as a method to address the challenge of early detection in depression and depression-related mental health issues. Social media is a platform where individuals with and without mental health issues freely engage in conversations about both mental health and casual topics. There is prior work that uses social media posts for depression detection, and similarly, in this project we use Reddit and identify self-diagnosis disclosures of individuals with depression. We compare their messages across all subreddits with messages from control users and we build a few classifiers that attempt to identify depression in social media posts. Our best model is a supervised fastText model, which we use in an experiment to determine the weights of specific relevant keywords. Future work should tune parameters for better accuracy measures, attempt other word embedding techniques such as BERT, and should prioritize model sensitivity over specificity. Moreover, with the acquisition of the LIWC library, further analysis of the data can be conducted to improve the shape and nature of the dataset.

1 Introduction

Natural Language Processing (NLP) has been applied extensively to binary classification problems, often with real-world implications. Sentiment analysis, translations, and fake news detection are just some of the problems NLP is extensively used in solving. In this project, we address a similar problem - processing a person's use of language for the the detection of depression.

Mental illnesses can greatly impact the quality of life and wellness of individuals in society ((Strine et al., 2008),(Mowery et al., 2017)), with one of the most common and, crucially, most subtle illness being depression. Depression is one of the leading causes of the 800,000 deaths by suicide globally each year, making timely detection and intervention important. However, it is still associated with a stigma of mental illness and affected individuals may not seek help - making detection difficult. At this time, social media platforms such as Reddit and Twitter are an ever-growing location of conversation, where people often talk openly through a layer of anonymity. This presents an opportunity to use social media posts to detect depression for early intervention. Of course, this opportunity comes with the responsibility to protect and respect the privacy of individuals suffering from mental illnesses.

In this paper, we will be using social media posts to identify language patterns correlated with depression. This can help enrich our understanding of mental health illnesses and how to better conduct early detection.

2 Background and Related Work

An early demonstration of depression detection from social media posts was carried out by Coppersmith et. al. in 2014 using Twitter posts (Coppersmith et al., 2014). In addition to depression, they also trained models to detect three other types of types of mental illnesses: post-traumatic stress disorder, bipolar disorder, and seasonal affective disorder. (Coppersmith et al., 2014) used 'positive' and 'negative' n-gram language models on the posts. This seems appropriate for short, simple Twitter posts. A potential caveat to using short messages by users could come from sort of information Twitter posts convey - given Twitter's character limit, users might be posting in a more thoughtful manner than an unfiltered or honest one.

Later studies have tried various models and dataset collection materials. Relevant approaches include deep learning models, used in projects such as (Orabi et al., 2018), (Haque et al., 2021) or the Deep Bag-of-Sub-Emotions (DeepBoSE), a deep learning model which incorporates emotional information internally from (Lara et al., 2021). In (Cohan et al., 2018), the best model from a wide range of competitors was a supervised fastText model.

In 2017, the RSDD dataset was compiled by Yates et. al. with Reddit posts pertaining to depression, also on the basis of self-diagnosis. In 2018, SMHD[1] was released by Cohan et. al. as a resource containing a large social media dataset as well as phrases and patterns to identify self-diagnosis by users for nine different types of mental illnesses, including depression. The authors test the precision of these patterns in identifying users who are likely to be suffering from mental illnesses. The depression-related patterns and phrases from this resource were crucial in building our dataset for this project, as described in a later section.

With this background, the goal for our project is reimplementing a few models of depression detection in a user-centered manner. We aim to detect not only the comments on topics related to depression, but also comments on other topics that are made by depressed users. We will compare how differently trained models weigh specific words in classification, and how various word embeddings and model architectures impact performance.

One potential easier way to address our issue is to compare messages which are part of depression related conversations to messages that are part of casual conversation. This can be achieved through, for instance, categorizing whether a message was posted in a depression-related subforum or an unrelated subforum of the social media site. We believe that such an approach would fail to capture the crucial fact that depressed and non-depressed individuals engage with each other in all types of conversations: non-depressed social media users can use the platform to engage with depressed individuals, and depressed individuals both often engage in casual conversations and can avoid talking about their mental health issues on social media platforms. Therefore, we are interested in building a classifier that is less biased on the type of a given message, i.e. what kind of conversation it belongs to, and that cares more about the pattern of language used in the social media post which can reveal information about the poster.

Multiple past papers have taken into consideration that mental illnesses often appear as comorbidities - people may be suffering from more than one at a time. (Coppersmith et al., 2014), (Cohan et al., 2018). Models which try to classify multiple mental illnesses, as well as datasets that label for the presence of multiple types of mental illnesses take this into account. We have focused only on depression detection here since we believe even just working on this binary classification problem can provide insights into what works and what does not in analyzing the language used by depressed people. However, we do bear in mind that a multi-value classification problem is the ideal future consideration and application of whatever architecture is built for a simple depression detection model.

3 Data Collection

This study used a dataset of anonymized Reddit comments from a mix of depressed and non-depressed users. The choice of Reddit as a platform was based on the ease of scraping post information and the fact that this platform supports longer comments (unlike Twitter, which has a character limit) that can be more informative and more similar to natural language conversations. After running into an issue with the lack of easily available depression detection datasets, we built and preprocessed a dataset as the initial part of the project.

3.1 Comment Collection

Comments were obtained from specific Reddit 'subreddits' (forums) using the `praw`[2] package. Comments were first scraped from three depression-related subreddits: r/depression, r/depressed, and r/depression_help. With ethical considerations in mind, only public comments were obtained. Authors were anonymized by replacing each username with a unique numeric ID, without maintaining any record of the correspondence. These comments were then checked for self-diagnosis phrases such as 'My {doctor} diagnosed me with {depression}' or 'since my {depression} diagnosis'. To make sure that the disclosure comments were personal disclosures and not revealing of someone else's mental health status, we also

[1] cite this, https://arxiv.org/abs/1806.05258v2

[2] https://github.com/praw-dev/praw

used negative detection patterns ('{ref} was diagnosed') to filter the posts that initially passed the positive diagnosis test. The self-diagnosis phrases we used were from the SHMD paper ((Cohan et al., 2018)). Comment authors who had used at least one of these phrases in at least one post were added to a 'depressed authors' set. Their comments, unrelated to self-diagnosis, were then collected from a variety of subreddits. Comments from depressed authors were stored with labels of 1, which indicates a positive value for the binary question 'Is the author depressed?'

Separately, control authors not belonging to the 'depressed' set were chosen from casual conversation subreddits such as r/casual, r/chat, and r/AskReddit. The control authors' comments were scraped and stored with labels of 0. The final dataset we constructed had 82,346 comments from self-diagnosed depressed authors and 100,000 comments in the control set. Only comments with a minimum character length of 15 were selected.

3.2 Preprocessing

Comments scraped from Reddit had to be preprocessed and simplified for our model to effectively use them, and we took multiple steps to do so. Mixed-case comments were converted to lowercase. Emoji were replaced with their text equivalents using the `emoji` Python package, so as to conserve the information they contain. Comments with a character length lower than a minimum threshold (either 100 or 200, depending on the model) were removed. For some models, stemmers and lemmatizers from the nltk library were used. We also removed links, names of other subreddits, and the usernames of any users who may have been mentioned in the post. Comments that had been deleted or removed since posting were filtered from the dataset, as were auto-generated suicide helpline comments.

We also trained FastText word embeddings (unsupervised FastText) on the data to visualize the sort of words that were better represented in our dataset. Although we were not able to obtain a working model using these embeddings, the nearest neighbors of a few words give us a good idea of the subjects that a potential model would be better or worse at handling.

Word	CBOW Nearest Neighbors
'depressed'	'depressed,', 'anxious', 'repressed', 'suicidal,', 'impressed', 'depressed.', 'anxious,', 'suicidal', 'stressed', 'diagnosed,'
'yogurt'	'yourself,', 'yourself!', 'yourself?', 'yourself', 'u', 'yourself.', '*you*', 'youd', 'you,', 'youre'
'happy'	'happy!', 'happy?', 'happy,', 'unhappy', 'happy.', 'happier,', 'happily', 'hurtful', 'happier', 'unhappy.'
'sad'	'sad,', 'bad', 'cry,', 'guilt', 'ok', 'mad', 'angry', 'tell,', 'mad,', 'kinda'

Word	Skipgram Nearest Neighbors
'depressed'	'depressed,', 'depressed.', 'depressing', 'depressing.', 'depressive', 'depression?', 'depression', 'depression,', 'repressed', 'depression.'
'yogurt'	'snail', 'microwave', 'pH', 'pepper', 'bumper', 'TFS', 'ketchup', 'dryer', 'pikachu', 'bbq'
'happy'	'happy!', 'happy?', 'happy,', 'unhappy', 'happy.', 'happier,', 'happier', 'happily', 'unhappy.', 'happiest'
'sad'	'sad.', 'sad,', 'hopeless', 'ungrateful', 'ungrateful,', 'lonely,', 'hopeless.', 'misery.', 'unbelievable', 'hurtful'

An unsurprising observation is multiple word representations for a single word with varied punctuation. Whether or not to strip punctuation (or separate it from the words) is similar to deciding whether or not to preserve capitalization. It conveys information, but adds complexity, increasing the vocabulary significantly. In our dataset, we prioritized the information contained in punctuation, including the information it adds to the word itself (in some cases, being used even as a morpheme to add emphasis or an emotion).

It is also expected that given the large proportion of depressed users in our comment authors, the embeddings are better for words that convey emotions than random words such as 'yogurt'. The distinction between the nearest neighbors for 'happy' and 'sad' is an interesting one. We see many morphological variants for 'happy', and semantically related but morphologically distinct words as neighbors

of ‘sad’. This may be because ‘sad’ and words conveying negative emotions are used more often in the dataset than words containing positive emotions, and the negative emotion words might be often used together or in similar contexts. Another possibility, though less likely, is that there are simply more variations on the word ‘happy’, perhaps due to its various morphological forms or the larger variety of punctuation that can be used with it.

After preprocessing and analyzing the dataset, we constructed classification models, assessed their accuracy, and carried out experiments using their predictions.

4 Methodology

4.1 Analysis

Although it was reported ((Cohan et al., 2018)) that Redditors (self-)diagnosed with depression-related conditions write, on average, messages roughly twice as long as control Redditors, the difference / ratio between average self-diagnosed (SD) message length (134) and average control length (112) is much lower.

Figure 1 shows visual representations of the sub-datasets grouped by label. Some of the differences between word clouds are unsurprising, such as the increased frequency of the word ‘feel’ in the S.D. dataset, and some are a bit surprising, such as the increased frequency of the word ‘really’ in the S.D. dataset.

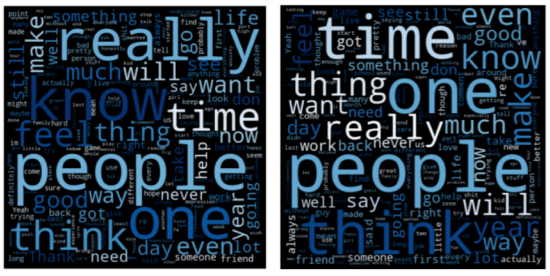


Figure 1: Visual Representations for messages written by users self-diagnosed with depression (Left Word Cloud) and by control users (Right Word Cloud).

4.2 Models

In order to devise a successful message classifier targeted for depression detection, we explored several methods (see Figure 2). For each, we trained our classifiers on a binary label setting, on both messages which are and which are not associated with any condition related to depression. The methods we considered are the following:

Model	Post-length-min	Eval	Eval-value
RandomForestClassifier	100	AUC_test	0.63
Supervised Fasttext	200	precision	0.644
LSTM (Glove)	200	training/val accuracy	0.6/0.58

Figure 2: A summary of the best performances for each model attempted

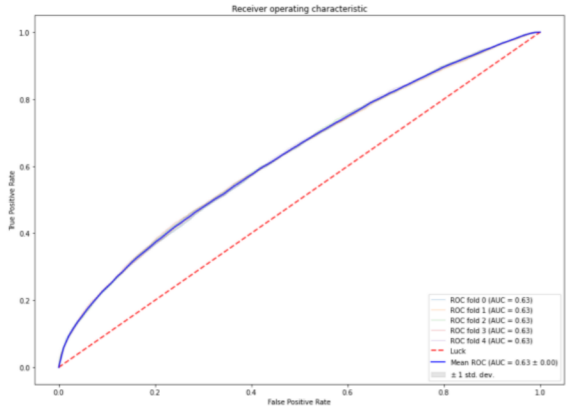
Random Forest Classifier A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. We trained a random forest classifier with bootstrapping and 1000 estimators. The input text was pre-processed with a stemmer and a lemmatizer. A K-Fold cross validation was used to avoid overfitting (K=5). Figure 3 shows the receiver operating characteristic (ROC) curve that shows the classifier’s performance at each fold followed by the Area Under the Curve (AUC) scores. The model is thus successful at discriminating positive and negative labels 63% of the time.

Long short-term memory (LSTM) Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture with feedback connections. We trained a bidirectional LSTM with Glove word embeddings. The most successful attempt used the embedding dimension of 16, with input length of 100 and was trained for 7 epochs. It lead to 60% traning and 58% validation accuracy. The accuracies and losses are shown in Figures 4 and 5, respectively.

Supervised FastText : FastText[3] is a library for efficient learning of word representations and text classification. FastText supports supervised (classifications) and unsupervised (embedding) representations of words and sentences. We used supervised FastText for our classifier. We trained for 19 epochs with a learning rate of 0.1 and 3 word ngrams. We obtain 64.4% precision. Out of all the models attempted, the fastest and most accurate model is Supervised FastText, which is also elicited the best performance when investigated by (Cohan et al., 2018).

[3]https://www.analyticsvidhya.com/blog/2017/07/word-representations-text-classification-using-fasttext-nlp-facebook/

While the dataset was constructed by only considering messages longer than 15 characters, this still invites many potentially meaningless messages where the effect of depression-related conditions is hard to be argued. To avoid feeding the model a disproportionately large amount of potentially noisy input, we decided on to pose a minimum message length restriction on the dataset each time a model was trained. The message length restrictions for each model are included in Figure 2.



	AUC Train	AUC Test
0	0.999996	0.627313
1	0.999998	0.626161
2	0.999979	0.631820
3	0.999989	0.631564
4	0.999989	0.630374

Figure 3: ROC curve and AUC scores for the Random Forest Classifier

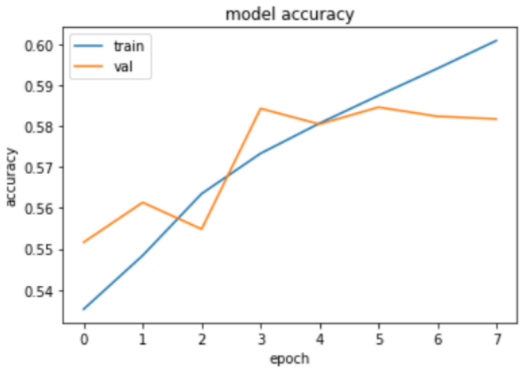


Figure 4: Accuracy for LSTM with Glove word embeddings.

4.3 Experiments

Besides designing a model successful at correctly classifying user depression from a given message, we were also interested in investigating the particular weights the model set on specific

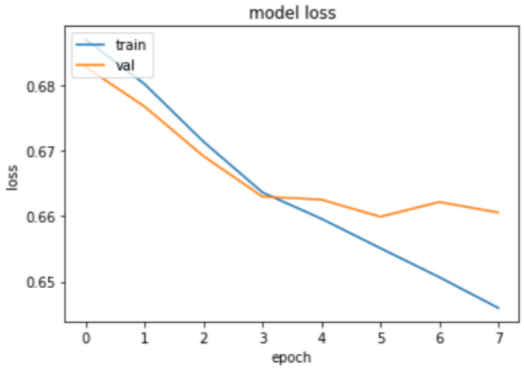


Figure 5: Loss for LSTM with Glove word embeddings.

keywords. Thus, driven by the visual representations from Figure 1 and by world knowledge about depression-related conditions, we designed the following experiment: We compiled a list of short 0-labeled messages and a list of specific keywords we suspected would have a large weight in our model, we systematically created natural sentences by merging the keyword with each given sentence, and we recorded how often each keyword changes the label of a given message from 'not depressed' to 'likely depressed'.

Table 1 shows the 10 keywords chosen and the scores tested on 50 short sentences. The words 'depress' and 'therapist' have near-perfect score, which is expected. From the visual representations of the data, we also expect high score for the words 'feel', 'life' and 'really'. Also from the visual representations, 'know' and 'think', which are almost interchangeable in natural language, have complementary distributions which translate into the experimental scores. There also seem to be an unexpected gender bias: sentences prefaced with 'As a man' are much more likely to be labeled as depressed as opposed to the same sentence prefaced with 'As a woman'. This might be due to a larger male population in the subredditor group, or just due to the higher frequency of the word 'man' which can sometimes be used genderlessly. An example of how the keywords are merged with a given sentence is given in Table 2. Not all words that we expected to be correlated to depression related mental health issues generate a high score in the label-changing experiment. For instance, the keyword 'sad', although our intuition correlates it highly with depression, perhaps due to its high frequency, scores fairly poorly in our experi-

Keyword	Score
feel	0.68
sad	0.28
know	0.30
think	0.02
really	0.36
man	0.34
woman	0.02
life	0.64
depress	0.90
therapist	0.96

Table 1: Label changing scores for specific keywords.

I feel like you are selfish
You are selfish, which is sad
I know you are selfish
You are really selfish
As a man, I think you are selfish
All my life I've thought you are selfish
You are selfish, which is depressing
My therapist says that you are selfish

Table 2: The posivitely labeled keyword-modified sentences generated from the sentence 'You are selfish.'

ment. The mystery beyond the word 'really' could be explained in a couple of way: it is a frequent filler word which could be an indicator of higher message length, previously correlated with mental illnesses (Cohan et al., 2018); or it is commonly used to enhance feeling statements, and we know from both the visual representations and our experimental scores that messages from depressed individual frequently convey feelings.

5 Conclusion and Future Considerations

Our best model was a supervised FastText model with 64.4% precision. However, accuracy measures may not be the absolute best way to assess the performance of a model designed for depression detection. In a situation where early intervention is necessary, a good model would rather misclassify a 'non-depressed' message as 'depressed', than miss to identify a 'depressed' message. False positives are thus less harmful to applications of a depression detection model than false negatives are, and a near future improvement to our model would be fine tuning it to prioritize better sensitivity over specificity.

In terms of other methodological approaches, we would like to tweak both model architecture and word embeddings in the future. One of the models we described here used GloVe embeddings in an LSTM architecture. GloVe embeddings are trained on the co-occurrences of entire words and might miss out on essential morpheme-level information as well as word derivatives and misspellings. Additionally, pre-trained GloVe embeddings were not specific to our dataset. The unsupervised FastText embeddings described in the preprocessing section address these issues. They are trained on the unlabeled dataset rather than a generic text such as the English Wikipedia, and can represent the finer differences between similar, depression-related words. FastText is also trained on character-level n-grams, making it much more ideal for capturing morpheme-level information and derivatives.

Word embeddings that are trained on a larger corpus do, however, have the advantage of likely being more accurate. From this perspective, we are interested in the result of using embeddings from BERT or fine-tuning DistilBERT to perform this particular classification task. We anticipate that this runs the risk of not accounting for social-media-specific slang, punctuation differences, or misspellings. The question is whether such an effect exists, and if it does, whether fine-tuning on our Reddit dataset can reduce the effect to the point where this is still a highly useful model.

As discussed in the related work section, model architectures other than the ones implemented here have been used by researchers in the past. A particularly promising one seems to be a Convolutional Neural Network (CNN) model (Rodrigues Makiuchi et al., 2019). An advantage of CNNs over RNNs such as LSTMs is that they do not depend on the computations of the previous time step in sequence processing. Thus,implementing this model architecture could also be a future consideration.

The 'Linguistic Inquiry and Word Count'[4] (LIWC) analyzes texts and calculates how people use several different word categories across a wide range of texts. LIWC allows for determination of emotions (positive or negative), self-references, long words or words that refer to specific subjects. LIWC was designed to quickly analyze over 70 language dimensions through hundreds of text samples in seconds. Using this program is locked be-

[4]https://liwc.wpengine.com/

hind a license, but doing so could provide crucial insight into the nature of our dataset and might provide useful guidelines into how to change or how to adapt our models. LIWC was used in several pieces of related work, including (Cohan et al., 2018), (Mustafa et al., 2020) and (Islam et al., 2018).

References

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. *arXiv preprint arXiv:2102.09427*.

Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6(1):1–12.

Juan S Lara, Mario Ezra Aragon, Fabio A Gonzalez, and Manuel Montes-y Gomez. 2021. Deep bag-of-sub-emotions for depression detection in social media. *arXiv preprint arXiv:2103.01334*.

Danielle Mowery, Craig Bryan, and Mike Conway. 2017. Feature studies to inform the classification of depressive symptoms from twitter data for population health. *arXiv preprint arXiv:1701.08229*.

Raza Ul Mustafa, Noman Ashraf, Fahad Shabbir Ahmed, Javed Ferzund, Basit Shahzad, and Alexander Gelbukh. 2020. A multiclass depression detection in social media based on sentiment analysis. In *17th International Conference on Information Technology–New Generations (ITNG 2020)*, pages 659–662. Springer.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. 2019. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 55–63.

Tara W Strine, Ali H Mokdad, Lina S Balluz, Olinda Gonzalez, Raquel Crider, Joyce T Berry, and Kurt Kroenke. 2008. Depression and anxiety in the united states: findings from the 2006 behavioral risk factor surveillance system. *Psychiatric services*, 59(12):1383–1390.